AD_______________

GRANT NUMBER DAMD17-94-J-4237

TITLE: Development of a Stochastic Simulation Model of the Cost-Effectiveness of Promoting Breast Cancer Screening

PRINCIPAL INVESTIGATOR: Nicole Urban, Ph.D.

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research
Seattle, Washington 98104-2092

REPORT DATE: September 1996

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 1996

3. REPORT TYPE AND DATES COVERED
Annual (22 Aug 95 - 21 Aug 96)

4. TITLE AND SUBTITLE Development of a Stochastic Simulation Model of the Cost-Effectiveness of Promoting Breast Cancer Screening

5. FUNDING NUMBERS
DAMD17-94-J-4237

6. AUTHOR(S)
Nicole Urban, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fred Hutchinson Cancer Research
Seattle, Washington 98104-2092

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

Year 02:

During the second year of this project, the components of the model of breast cancer screening were outlined and coding begun. The literature on the natural history of breast cancer was carefully reviewed with consulting experts, and a modeling strategy was developed based on the findings. Model code is divided into independent modules to facilitate future changes and enhancement. A collaboration with researchers at NCI was initiated to extend the model to defined populations, and a paper was submitted for publication.

14. SUBJECT TERMS Mammography, Screening, Modeling, Cost Effectiveness, Promotion, Simulation, Humans, Data, Breast Cancer

15. NUMBER OF PAGES
37

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Nicole Urban 2/6/97

PI - Signature Date

Annual Report for Grant DAMD17-94-J-4237

August 22, 1995 - August 21, 1996
Year 02

Development of a Stochastic Model of the Cost-Effectiveness of Promoting Breast Cancer Screening

Nicole Urban, ScD
Principal Investigator

Table of Contents

Introduction

The purpose of this four-year project, funded in August 1994, is to identify an efficient strategy for reducing breast cancer mortality through breast cancer screening. To identify such a strategy, the trade-off between the frequency of screening among participants and the promotion of participation among underusers will be investigated. Ways to improve the effectiveness of screening in women aged 40-49 will be investigated, using new biomarkers and detection modalities, and the relative cost-effectiveness of various interventions to promote the use of regular breast cancer screening among women aged 50-80 will be investigated. A comprehensive stochastic simulation model of the effectiveness and cost-effectiveness of breast cancer screening will be developed, and its key parameters estimated.

Body

Year 02 was spent on identifying and answering the specific questions needed to understand how breast cancer affects a defined population. A Statistical Research Associate, Matt Gable, was hired in January 1996 to analyze the problem and create the specific algorithms and parameter sets required by the model. Mr. Gable replaced Chris Colby (50%) and Adelina Tseng (30%) in the original proposal. Much effort was given to the review of the clinical aspects of breast cancer, including tumor growth, ductal carcinoma in situ (DCIS), staging, and screen test characteristics. Many key features of the ovarian cancer model were extracted to be used in the breast cancer model, which was substantially reconceptualized. An implementation of the model was begun, designed around modules for natural history (including population characteristics and disease progression), screening, survival, and costs and benefits . Each module will be programmed separately and all will work together to create the complete simulation model.

Dr. Urban and Mr. Gable participated in the National Cancer Institute's (NCI) meeting with investigators from the Netherlands on their colon cancer modeling project, MISCAN, in Bethesda, Maryland during the first week of June. Dr. Etzioni was unable to attend due to the birth of her first baby in late April. This meeting gave investigators an opportunity to discuss stochastic simulation procedures and to exchange ideas and solutions for overcoming programming obstacles and other issues. Investigators at NCI invited Fred Hutchinson researchers to participate in a recently proposed project entitled POPSIM that is designed to incorporate existing cohort-based microsimulation models into a simulation engine capable of simulating multi-cohort populations. This work builds on previous activities by NCI investigators, including CAN*TROL and will include our ovarian and breast cancer simulation models and the prostate cancer screening model under development by Dr. Etzioni. Researchers from Fred Hutchinson, including Matt Gable, the programmer of the breast cancer model, are working in collaboration with POPSIM investigators.

A paper entitled "Estimating medical costs from incomplete follow-up data" has been submitted to and accepted by Biometrics (see Appendix). Dr. Etzioni enlisted the assistance of Danyu Lin from the Department of Biostatistics at the University of Washington and Eric Feuer of NCI in solving methodologic problems associated with estimation of attributable lifetime costs from censored cost data such as those in the SEER-Medicare file. This paper describes the methods developed by these collaborating statisticians to analyze the SEER-Medicare cost data to estimate expected lifetime costs attributable to stage-specific breast cancer. We are in the process of obtaining updated SEER-Medicare data from the Health Care Financing Administration in response to the request which was submitted in Year 01.

Details of each aspect of the work completed in Year 02 are described below.

Natural History Module

This module generates characteristics of interest for a cohort of women, including presence/absence of breast cancer, age at diagnosis, type of tumor (DCIS, invasive, or metastatic), size and growth rate (or doubling time), and ages at metastasis, invasion, and onset. Changes in breast density are assigned, as well as age at death from competing mortality in the absence of cancer.

The natural history module is size-driven: most events in disease progression are associated with a tumor size, and time of an event is calculated via the tumor doubling time. The module also runs backwards: the initial event in the model is clinical detection, and earlier events are generated in order backwards to onset.

Screening Module

This module produces a screening schedule for each woman using a method adopted from the MISCAN model, then determines results. The MISCAN method of scheduling allows for many types of schedules, including regular intervals, irregular "naturalistic" participation, and single use, and does not require 100% participation. Results are calculated using the sensitivity and specificity of screen tests, modified by breast density and size of tumor.

Survival Module

Survival after diagnosis of breast cancer is calculated using Kaplan-Meier survival curves, as planned by Ruth Etzioni. Age and stage at diagnosis determine the curve to be used, and the same set of curves is used for both clinical and screen diagnosis. Survival is added to the relevant age at diagnosis to obtain age at death. If desired, survival may be linked to the aggressiveness of the tumor; this is accomplished by using the same random number to select both tumor doubling time and survival.

Costs & Benefits Module

In this module, costs associated with screening and treatment are discounted and summed. Years of life saved (YLS) are calculated as the difference in age at death after clinical and screen detection. A summary performance measure, cost/YLS, is calculated for the screening program.

POPSIM

NCI researchers Eric Feuer and Julie Legler proposed, and we agreed to, a collaboration to use our models as the engines in a larger population model. The ovarian and breast cancer models are built to simulate a single cohort of women at a time, but some projects—for example, projecting trends in the US population—require simulation of multiple cohorts. The POPSIM project will add to the existing models the capability to update parameters automatically from cohort to cohort and aggregate the results for a defined population.

Conclusion

Summary of Year 02

During the second year of this project, the design phase of the modeling effort was completed. The components of the model of breast cancer screening were outlined, the relationships among them were specified, and coding was begun. The literature on the natural history of breast cancer was carefully reviewed in collaboration with breast cancer experts and a modeling strategy was developed based on the findings. Model code is divided into independent modules to facilitate future changes and enhancement. A collaboration with researchers at NCI was initiated to extend the model to defined populations, and an on-going communication via e-mail was established with investigators in the Netherlands following a meeting about the MISCAN modeling work. Model development work has progressed in accordance with the original timeline, but analysis of cost data is behind schedule due to delays in receiving data from HCFA. Methodologic work on cost data analysis proceeded in the absence of data, resulting in a paper which is forthcoming in Biometrics.

Plans for Year 03

The next year of the project will see the completion of code for the model, derivation of parameter sets, and integration with a graphical Basic User Interface being developed for it and the ovarian model. Parameter sets will be derived from 1) NCI's SEER database, for survival and stage- and age-specific incidence in absence of screening; 2) Dr. Urban's Mammography Quality Improvement Project (MQIP), for the effects of breast density on the risk of an interval cancer; 3) Dr. Urban's Community Mammography Trial (CMT) for women's participation in screening; and 4) the SEER-Medicare database for treatment costs as a function of stage at diagnosis and time since diagnosis.

Design questions yet to be resolved include the precise definitions of stages for breast cancer, inclusion of cancers never diagnosed in the absence of screening, and modeling changes in breast density. Model testing and validation will begin subsequent to the resolution of these questions and code completion.

A network of advisors is being formed to provide additional input. This network currently includes Mariann Drucker, a radiologist at the University of Washington, and Ben Anderson, a surgeon specializing in breast cancer, also at UW. The advisors are consulted on key points in the model design and provide critical review of working documents related to the model.

A full report, containing more details on the workings of the model, is due in May 1997 to the DoD and is currently being prepared. This report describes the assumptions made in construction of the model, techniques used, options available to the user, and strengths and weaknesses of the model. If available by May 1997, preliminary results may also be described.

Appendix A

Estimating Medical Costs from Incomplete Follow-up Data

Submitted Manuscript
Do Not Cite

Estimating Medical Costs from Incomplete Follow-up Data

D. Y. Lin

Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA

E. J. Feuer

Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD 20892, USA

R. Etzioni

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA

and Y. Wax*

Department of Statistics, Hebrew University, Jerusalem 91905, Israel

Summary

Estimation of the average total cost for treating patients with a particular disease is often complicated by the fact that the survival times are censored on some study subjects and their subsequent costs are unknown. The naive sample average of the observed costs from all study subjects or from the uncensored cases only can be severely biased, and the standard survival analysis techniques are not applicable. To minimize the bias induced by censoring, we partition the entire time period of interest into a number of small intervals, and estimate the average total cost either by the sum of the Kaplan-Meier estimator for the probability of dying in each interval multiplied by the sample mean of the total costs from the observed deaths in that interval or by the sum of the Kaplan-Meier estimator for the survival probability at the start of each interval multiplied by an appropriate estimator for the average cost over the interval conditional on surviving to the start of the interval. The resultant estimators are consistent if censoring occurs solely at the boundaries of the intervals. In addition, the estimators are asymptotically normal with easily estimated variances. Extensive numerical studies show that the asymptotic approximations are adequate for practical use and the biases of the proposed estimators are small even when censoring may occur in the interiors of the intervals. An ovarian cancer study is provided.

Key words: Censoring; Cost analysis; Economic evaluation; Health services; Medical care; Missing data; Resource utilization; Survival analysis; Treatment cost.

* Y. Wax is deceased. This paper is dedicated to his memory.

1. Introduction

Recent years have seen a heightened interest in studying the cost of health care. One important component of this effort is the assessment of medical costs for treating a disease. For instance, when evaluating the cost-effectiveness of a cancer screening program, the potential savings in treatment costs due to earlier diagnosis through screening are of interest. As another example, comparisons of the average costs associated with alternative therapies may lead to substantial cost reduction. The data for such analysis may be derived from clinical trials, disease registries, health insurance records, etc. A common feature with the available data sources is that some patients are not followed for the entire durations of interest. (If the τ-year cost is under study, then the duration of interest is the minimum of τ and the patient's survival time.) Thus, the durations of interest on these patients are censored and their subsequent costs are unknown.

Until recently, the average total cost for a group of patients has commonly been estimated by the sample mean of the observed costs from all study subjects or from only the uncensored cases. The former estimator, to be referred to as the full-sample estimator, is always biased downwards since the costs incurred after censoring times are not accounted for. The latter estimator, called the uncensored-cases estimator, is also destined to be biased: it is biased towards the costs of the patients with shorter survival times because larger survival times are more likely to be censored. The difficulties with using such naive sample averages to make inference about the survival distribution have long been recognized in the field of survival analysis.

In an attempt to adjust for the effects of censoring, several researchers (e.g., Quesenberry et al., 1989; Hiatt et al., 1990; Fenn et al., 1995) have applied the standard survival analysis techniques (e.g., Kaplan-Meier estimator and log rank test) to the problem of cost evaluation by treating costs as potentially right-censored survival times (i.e., attaching the censoring indicator to the observed total cost). This strategy, however, is invalid unless all patients accumulate costs with a common (deterministic) rate function over time (yielding a one-to-one correspondence between the survival time and total cost). In practice, the cost functions vary among patients. Thus, a patient who accrues costs at higher rates tends to generate larger total costs at both the survival time and censoring time, which implies that the total cost at the survival time is positively correlated with the total cost at the censoring time. This correlation implies that "censored" total costs cannot be analyzed by standard survival analysis methods, all of which require independence between the variable of interest and its censoring variable.

In this paper, we show how to properly adjust for censoring in the cost estimation. Specifically, we divide the entire time period of interest into several intervals and then estimate the average total cost by the sum of the Kaplan-Meier estimator for the probability of dying in each time interval multiplied by the sample mean of the total costs from those who are observed to die in that interval. The distribution function of the total cost and its quantiles may be estimated in a similar fashion. If the costs accrued within the intervals are recorded, then the average total cost can also be estimated by the sum of the Kaplan-Meier estimator for the probability of surviving to the start of each interval multiplied by an appropriate estimator for the average cost over the interval conditional on surviving to the start of the interval. The latter approach makes fuller use of the cost information and accommodates left-truncation in addition to right-censorship. Consistent estimators can be constructed under either approach if the censoring time distribution is discrete. Furthermore, the proposed estimators are asymptotically normal with variances that can be easily estimated. Extensive numerical studies show that the new methodology is appropriate for practical use. A detailed illustration with an ovarian cancer study is provided.

2. Methods

2.1. Preliminaries

Let the random variable C denote the total medical cost for a patient over the time period $[0, \tau)$. In addition, let T and U be the latent survival and censoring times. We assume that T is continuous and U is either continuous or discrete. If $T < \tau$, then C becomes the total cost up to T. Our main task is to estimate the mean (total) cost $E = \mathcal{E}(C)$, where $\mathcal{E}$ denotes expectation. If no patient is followed beyond τ, then it would not be possible to include the cost incurred after τ in the definition of the mean total cost without imposing stringent and untestable assumptions.

We divide the entire time period $[0, \tau)$ into K intervals $[a_k, a_{k+1})$ $(k = 1, \ldots, K)$, where $a_1 = 0$ and $a_{K+1} = \tau$. Let the random variable C_k be the cost incurred over $[a_k, a_{k+1})$ $(k = 1, \ldots, K)$. Naturally, a patient can accrue costs over $[a_k, a_{k+1})$ if and only if he/she survives to the start of the interval a_k.

Write $X = \min(T, U)$ and $\delta = I(T \leq U)$, where $I(\cdot)$ is the indicator function. The data typically consist of n independent replicates of $(X, \delta, \tilde{C})$, where $\tilde{C}$ is the observed total cost, namely, the cost accrued from the start of the follow-up to the last contact date X. When the cost histories are recorded, $\tilde{C}$ may be decomposed as $(\tilde{C}_1, \ldots, \tilde{C}_K)$, where $\tilde{C}_k$ is the *observed* cost over $[a_k, a_{k+1})$. The subscript i will be added to the variable names to indicate individual patients. Obviously, $\delta_i = 1$ or $X_i = \tau$ implies that $\tilde{C}_i = C_i$. If $X_i < a_k$, then $\tilde{C}_{ki}$ is either zero or missing dependent on whether $\delta_i = 1$ or 0. Given $X_i \geq a_k$, $\tilde{C}_{ki} = C_{ki}$ if the ith patient is not censored before a_{k+1} and $\tilde{C}_{ki}$ equals the cost accrued over $[a_k, U_i)$ otherwise. We allow some of the $\tilde{C}_{ki}$ to be missing before X_i, but only in a completely random fashion.

The validity of the familiar survival analysis techniques, such as the Kaplan-Meier estimator, depends critically on the assumption of independent censoring, which requires that, at any follow-up time t, patients cannot be censored because they are at unusually high (or low) risk of dying (Kalbfleisch and Prentice, 1980, pp. 40–41). In our setting, it is necessary to extend this definition to require that, at any follow-up time t, patients are not censored because they will accrue unusually high (or low) costs.

We shall provide two approaches to cost estimation, one requiring only the observed total costs at the last contact dates and one making use of the observed costs within the small intervals $[a_k, a_{k+1})$ $(k = 1, \ldots, K)$. We begin with the latter approach.

2.2. Using the Cost Histories

Recall that $C = \sum_{k=1}^{K} C_k$, which implies that $E = \sum_{k=1}^{K} \mathcal{E}(C_k)$. By the conditional expectation argument, $E = \sum_{k=1}^{K} \mathcal{E}\left\{\mathcal{E}(C_k | T \geq a_k)\right\}$, which equals $\sum_{k=1}^{K} \Pr(T \geq a_k)\mathcal{E}(C_k|T \geq a_k)$, or

$$E = \sum_{k=1}^{K} S_k E_k,$$

where $S_k = \Pr(T \geq a_k)$ and $E_k = \mathcal{E}(C_k | T \geq a_k)$. The replacements of S_k and E_k $(k = 1, \ldots, K)$ by their consistent estimators will yield a consistent estimator for E. As will become clearer later, estimation of E_k $(k = 1, \ldots, K)$ is much less challenging than direct estimation of E.

The survival probabilities S_k $(k = 1, \ldots, K)$ can be consistently estimated by the Kaplan-Meier method. Specifically, let $t_1 < t_2 < \ldots < t_J$ be the (ordered distinct) observed survival times, and let d_j be the number of observed deaths at t_j and let n_j be the number of patients under observation at t_j $(j = 1, \ldots, J)$. Then the Kaplan-Meier estimator of S_k is

$$\hat{S}_k = \prod_{j: t_j < a_k} \frac{n_j - d_j}{n_j}.$$

The (extended) assumption of independent censoring mentioned in §2.1 implies that $\mathcal{E}(C_k|T \geq a_k) = \mathcal{E}(C_k|X \geq a_k)$ $(k = 1, \ldots, K)$, which enables us to estimate E_k from the patients who are under observation at time a_k. Let Y_{ki} indicate, by the values 1 vs. 0, whether or not $\tilde{C}_{ki}$ is included in the estimation of E_k. (Naturally, Y_{ki} is 0 if $\tilde{C}_{ki}$ is missing). Then E_k are estimated by

$$\hat{E}_k = \frac{\sum_{i=1}^n Y_{ki}\tilde{C}_{ki}}{\sum_{i=1}^n Y_{ki}}, \quad k = 1, \ldots, K,$$

and E is estimated by $\hat{E} = \sum_{k=1}^K \hat{S}_k \hat{E}_k$.

If we define $Y_{ki} = I(X_i \geq a_k)$, then $\hat{E}_k$ is the sample average of the *observed* costs over $[a_k, a_{k+1})$ among those who are under observation at the start of the interval. Denote the resulting estimator of E by $\hat{E}_A$. This estimator was previously utilized in some applications (Manning et al., 1989; Keeler et al., 1989; Hodgson, 1992; Etzioni, Urban and Baker, 1995), but its properties were not studied. It is easy to see that $\hat{E}_k$ is an unbiased estimator of E_k if censoring occurs only at the end of the interval, in which case $\tilde{C}_{ki} = C_{ki}$ for all non-zero Y_{ki}'s. This censoring pattern can be constructed if patients enter the study at discrete time points and are withdrawn from the study prematurely at limited time points. If censoring occurs before the end of the interval, then $\hat{E}_k$ will underestimate E_k since the costs from the censoring times to the end of the interval are not accounted for. Clearly, the bias of $\hat{E}_A$ depends on the amount and timing of censoring and diminishes as the intervals shrink. If there is heavy censoring and the cost information is available only in broad time intervals, then it may be advisable to prorate the costs of the censored cases.

It is instructive to compare $\hat{E}_A$ with the naive full-sample estimator mentioned in §1. (The latter may be regarded as an extreme special case of the former with $K = 1$.) For the naive estimator, the costs from the censoring times through the terminal time point τ are omitted entirely, which will result in substantial underestimation of E unless all the censoring times are close to τ. Even though we use the same kind of sample averages to estimate the individual E_k's, the resulting estimator $\hat{E}_A$ is always less biased because a censored case poses a difficulty only in the time interval in which the censoring occurs and the gap between the censoring time and the end of the corresponding interval is generally much smaller than that between the censoring time and the terminal time point τ.

An alternative way of estimating E_k is to exclude those who are censored during $[a_k, a_{k+1})$ from the calculation of the sample average $\hat{E}_k$. The resulting estimator will be unbiased if all the patients who are under observation at time a_k have the same probability of being censored during

$[a_k, a_{k+1})$. This condition, which guarantees that the uncensored C_{ki}'s are representative of all the C_{ki}'s in the kth interval, essentially requires that censoring occurs only at the start of the interval. (Larger survival times will have higher probabilities of being censored unless censoring takes place before any death.) Because the costs incurred over a small time interval are (stochastically) similar between the censored and uncensored cases, the sample average of the costs from the patients who are under observation at time a_k and who are not censored in $[a_k, a_{k+1})$ provides a reasonable estimator of E_k even when censoring may occur in the interior of the interval. The corresponding estimator of E, denoted by $\hat{E}_B$, is always less biased than the naive uncensored-cases estimator mentioned in §1, which may be regarded as an extreme special case of $\hat{E}_B$ with $K = 1$. The bias of $\hat{E}_B$ will approach zero as the interval widths decrease. In general, $\hat{E}_B$ is less biased than $\hat{E}_A$ but may also be less efficient. Naturally, both estimators reduce to the usual sample mean in the absence of censoring.

To summarize, $\hat{E}_A$ and $\hat{E}_B$ are consistent for E when censoring occurs only at the ends and the beginnings of the intervals, respectively, and both estimators are nearly consistent for small time intervals regardless of the censoring pattern. For making formal statistical inference about the average cost E, it is imperative to ascertain the distribution of the estimator $\hat{E}$ ($\hat{E}_A$ or $\hat{E}_B$). We show in Appendix A that, for large n, the estimator $\hat{E}$ is approximately normal with mean $E^* = \sum_{k=1}^{K} S_k E_k^*$ and with variance $\hat{V} = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} W_{ki} W_{li}$, where $E_k^* = \mathcal{E}(\tilde{C}_{k1} | Y_{k1} = 1)$, and

$$W_{ki} = \frac{\hat{S}_k Y_{ki} (\tilde{C}_{ki} - \hat{E}_k)}{\sum_{j=1}^{n} Y_{kj}} - \hat{S}_k \hat{E}_k \left\{ \frac{I(X_i \leq a_k)\delta_i}{R_i} - \sum_{j: X_j \leq \min(a_k, X_i)} \frac{\delta_j}{R_j^2} \right\},$$

and $R_i = \sum_{l=1}^{n} I(X_l \geq X_i)$. (As mentioned above, the use of $\hat{E}_A$ and $\hat{E}_B$ entails that $E^* = E$ if censoring occurs only at the ends and the beginnings of the intervals, respectively.) These results enable one to make formal inference about the average total cost for a group of patients or the difference between two independent groups.

2.3. Not Using the Cost Histories

By the law of conditional expectation, $E = \sum_{k=1}^{K} \mathcal{E}(C | a_k \leq T < a_{k+1}) \Pr(a_k \leq T < a_{k+1}) + \mathcal{E}(C | T \geq \tau) \Pr(T \geq \tau)$, or $E = \sum_{k=1}^{K+1} \mathcal{E}(C | a_k \leq T < a_{k+1}) \Pr(a_k \leq T < a_{k+1})$ with $a_{K+2} = \infty$.

Thus,

$$E = \sum_{k=1}^{K+1} A_k(S_k - S_{k+1}), \tag{2.1}$$

where $A_k = \mathcal{E}(C|a_k \leq T < a_{k+1})$.

Under the (extended) independent censoring condition given in §2.1,

$$A_k = \mathcal{E}(C|a_k \leq T < a_{k+1}, U \geq a_k) = \mathcal{E}(C|X \geq a_k, T < a_{k+1}), \tag{2.2}$$

showing that it is possible to estimate A_k from the patients who are under observation at time a_k. If censoring occurs only at the end of the interval (i.e., $a_{k+1}-$), then $I(X \geq a_k, T < a_{k+1}) = I(a_k \leq X < a_{k+1}, \delta = 1)$, in which case A_k can be consistently estimated by the sample mean of the total costs from those who are observed to die in $[a_k, a_{k+1})$, namely,

$$\hat{A}_k = \frac{\sum_{i=1}^n Y_{ki}\tilde{C}_i}{\sum_{i=1}^n Y_{ki}},$$

where $Y_{ki} = I(a_k \leq X_i < a_{k+1}, \delta_i = 1)$. This estimator may also be expressed as

$$\hat{A}_k = \frac{\sum_{i=1}^n Y_{ki}C_i}{\sum_{i=1}^n Y_{ki}}$$

since $Y_{ki} = 1$ implies that $\tilde{C}_i = C_i$. If censoring takes place at the start of the interval (i.e., a_k), then, given $\{X_i \geq a_k\}$, the T_i's have the same probability of being censored in the interval. Under this scenario, the patients who are observed to die in $[a_k, a_{k+1})$ are a random subset of all the deaths in $[a_k, a_{k+1})$, which entails that $\hat{A}_k$ is still consistent for A_k. If censoring occurs in the interior of the interval, then $\hat{A}_k$ tends to be driven by the costs of the patients who die early on in the interval because, given the same censoring distribution, larger survival times are more likely to be censored. However, if the interval is narrow, the costs associated with the early deaths of the interval are (stochastically) similar to those of the late deaths so that the bias of $\hat{A}_k$ will be small.

Most of the discussion in the preceding paragraph pertains to A_k $(k = 1, \ldots, K)$ only. By definition, $A_{K+1} = \mathcal{E}(C|T \geq \tau)$, which equals $\mathcal{E}(C|X \geq \tau)$ according to (2.2). Thus, we estimate A_{K+1} by

$$\hat{A}_{K+1} = \frac{\sum_{i=1}^n Y_{K+1,i}C_i}{\sum_{i=1}^n Y_{K+1,i}},$$

where $Y_{K+1,i} = I(X_i \geq \tau)$. Note that $\hat{A}_{K+1}$ is always consistent for A_{K+1} (regardless of the censoring pattern) provided that the independent censoring assumption holds.

It is clear from the expressions for $\hat{A}_k$ $(k = 1, \ldots, K+1)$ that the observed costs of the patients who are censored before τ are not involved in any calculations and therefore need not be recorded. The costs of the other patients (i.e., those who are observed to die or whose censoring times equal τ) are allowed to be missing, but only in a completely random fashion. Naturally, $Y_{ki} = 0$ if C_i is missing.

Given the $\hat{S}_k$'s and $\hat{A}_k$'s, we estimate E by

$$\hat{E}_T = \sum_{k=1}^{K+1} \hat{A}_k(\hat{S}_k - \hat{S}_{k+1}).$$

Because the Kaplan-Meier estimators and $\hat{A}_{K+1}$ are consistent regardless of the censoring pattern, the estimator $\hat{E}_T$ will be consistent as long as $\hat{A}_k$ $(k = 1, \ldots, K)$ are consistent. The consistency of the $\hat{A}_k$'s can be achieved if the censoring time distribution is discrete, in which case the cut-points a_k may be chosen to coincide with the possible censoring times. If the censoring distribution is continuous, then it is desirable to choose a fine partition of the time period so that the bias can be minimized. However, this is subject to the constraint that reliable estimation of A_k $(k = 1, \ldots, K)$ requires a reasonable number of observed deaths in each interval.

In order to study the large-sample properties of $\hat{E}_T$ under arbitrary censoring distributions, we define $A_k^* = \mathcal{E}(C|a_k \leq X < a_{k+1}, \delta = 1)$ $(k = 1, \ldots, K)$, $A_{K+1}^* = A_{K+1}$ and $E^* = \sum_{k=1}^{K+1} A_k^*(S_k - S_{k+1})$. Of course, if censoring takes place only at the boundaries of the intervals, then $A_k^* = A_k$ $(k = 1, \ldots, K)$ and $E^* = E$. We show in Appendix B that, for large n, the estimator $\hat{E}_T$ is approximately normal with mean E^* and with variance $\hat{V}_T = \sum_{i=1}^{n} \sum_{k=1}^{K+1} \sum_{l=1}^{K+1} W_{ki} W_{li}$, where

$$W_{ki} = \frac{(\hat{S}_k - \hat{S}_{k+1}) Y_{ki} (C_i - \hat{A}_k)}{\sum_{j=1}^{n} Y_{kj}} + \hat{E}_k(\hat{S}_{k+1} D_{k+1,i} - \hat{S}_k D_{ki}), \tag{2.3}$$

$$D_{ki} = \frac{I(X_i \leq a_k)\delta_i}{R_i} - \sum_{j: X_j \leq \min(a_k, X_i)} \frac{\delta_j}{R_j^2}, \tag{2.4}$$

and $R_i = \sum_{l=1}^{n} I(X_l \geq X_i)$.

In the special case where patients are enrolled simultaneously and study termination is the only source of censoring, the proposed estimator $\hat{E}_T$ reduces to the naive sample mean. As discussed in §1, when there are staggered patient entries and/or early withdrawals, the naive sample mean will be biased downwards if the observed total costs of all patients are included in the calculation and will be biased towards the costs associated with shorter survival times if the cases censored before τ are excluded. Although we use the sample means of the costs from the observed deaths to estimate the individual A_k's, the biases of the $\hat{A}_k$'s are minimal even if censoring may occur in the interiors of the intervals because the total costs of the patients who die in a small time interval are much more homogeneous than the total costs of all patients. Therefore, the bias of $\hat{E}_T$ is much smaller than the naive sample mean of the costs from all the observed deaths.

The idea of dividing the entire time period of interest into a number of small intervals and then combining the Kaplan-Meier estimators with the sample means is essential to the developments in both §2.2 and this subsection. The key difference lies in the decomposition of the costs. In §2.2, the mean total cost is decomposed as the sum of the probability of being alive at the start of each interval multiplied by the average cost incurred over the interval conditional on being alive at the start of the interval. Here, the same parameter is represented by the sum of the probability of dying in each interval multiplied by the average total cost of those who die in the interval. In the former case, the cost information on a patient contributes to the estimation of the E_k's for each time interval in which he/she is under observation, whereas in the latter case only the costs of the observed deaths are utilized. Some loss of efficiency may result from disregarding the costs of the censored cases. Additional difficulties may arise if the survival time is also subject to left-truncation (i.e., some patients have already accrued costs for some time before the follow-up begins). In that situation, the truncated cases have to be excluded from the calculation of the $\hat{A}_k$'s because the costs incurred before the truncation times are unknown. This may not only reduce the efficiency, but may also shorten the time period over which the average total cost can be estimated, as will be elaborated in §4. Of course, the main advantage of using $\hat{E}_T$ is that it does not require the breakdown of the medical costs in small time intervals, which is infeasible in some applications.

Another advantage of the approach taken in this subsection is that it can also be used to estimate the distribution function of C. Let $F(c) = \Pr(C \leq c)$ and $F_k(c) = \Pr(C \leq c | a_k \leq T < a_{k+1})$. Then, analogous to (2.1),

$$F(c) = \sum_{k=1}^{K+1} F_k(c)(S_k - S_{k+1}),$$

which can be (consistently) estimated by

$$\hat{F}(c) = \sum_{k=1}^{K+1} \hat{F}_k(c)(\hat{S}_k - \hat{S}_{k+1}),$$

where

$$\hat{F}_k(c) = \frac{\sum_{i=1}^n Y_{ki} I(C_i \leq c)}{\sum_{i=1}^n Y_{ki}}, \quad k = 1, \ldots, K+1.$$

By the arguments given in Appendix B, for large n, the estimator $\hat{F}(c)$ is approximately normal with mean $F(c)$ (or a limit analogous to E^* if censoring may occur in the interior of an interval) and with variance $\hat{V}(c) = \sum_{i=1}^{n} \sum_{k=1}^{K+1} \sum_{l=1}^{K+1} \tilde{W}_{ki}(c)\tilde{W}_{li}(c)$, where $\tilde{W}_{ki}(c)$ is the same as W_{ki} except that C_i and $\hat{A}_k$ on the right side of (2.3) are now replaced by $I(C_i \leq c)$ and $\hat{F}_k(c)$.

The quantiles (such as median) of the distribution can be estimated from the above empirical distribution function $\hat{F}$ along the lines of Brookmeyer and Crowley (1982). Let $0 < p < 1$, and define $\psi_p = F^{-1}(p) = \inf\{c : F(c) \geq p\}$ as the pth quantile of F. Then ψ_p is estimated (consistently) by

$$\hat{\psi}_p = \inf\{c : \hat{F}(c) \geq p\}.$$

An approximate 95% confidence interval for ψ_p is the collection of all values of ψ_p^0 which satisfy

$$\frac{|\hat{F}(\psi_p^0) - p|}{\hat{V}^{1/2}(\psi_p^0)} \leq 1.96,$$

i.e., all hypothesized values ψ_p^0 of ψ_p which are not rejected when the null hypothesis $\psi_p = \psi_p^0$ is tested against the alternative hypothesis $\psi_p \neq \psi_p^0$ at the 5% level based on the large-sample normality of $\hat{F}(c)$. This interval can be read directly from the upper and lower pointwise 95% confidence limits for $F(c)$ in the same manner as $\hat{\psi}_p$ can be read from the empirical distribution function $\hat{F}(c)$ itself, as will be demonstrated in §4.

3. Numerical Studies

Extensive Monte Carlo simulations were conducted to assess the performance of the proposed estimators $\hat{E}_A$, $\hat{E}_B$ and $\hat{E}_T$ in terms of bias, variance estimation and normal approximation. For comparison, we also evaluated the three naive estimators discussed in the previous sections. Let $\hat{E}_F$ and $\hat{E}_U$ denote, respectively, the naive full-sample and uncensored-cases estimators, and let $\hat{E}_{KM}$ denote the empirical mean of the Kaplan-Meier distribution treating total costs as right-censored survival times.

The survival times were generated from two distributions, uniform on $[0, 10)$ years and exponential with a mean of six years. We assumed that the study only lasts ten years so that the terminal time point τ equals 10. Thus, the average 10-year cost is the parameter of interest under both distributions.

We postulated U-shaped sample paths for the costs, i.e., there is a high cost associated with diagnostic tests around the time of diagnosis and then there is a sharp rise of costs prior to death. To be specific, we divided the entire time period into ten one-year intervals. Within each interval, there is a baseline cost which has a uniform distribution on $[1000, 3000]$ dollars annually; the total diagnostic cost is uniform on $[5000, 15000]$ dollars at the time of diagnosis; the cost in the final year of life is uniform on $[10000, 30000]$ dollars. (The diagnostic and death costs are added to the baseline costs.) We assumed that the annual baseline cost is evenly distributed over the one-year period and that the terminal-phase cost is evenly distributed in the final year of life (which may overlap two intervals). Given the above cost specifications, the mean costs over the ten-year period are about \$39000 and \$34680 under the uniform and exponential survival distributions, respectively.

The results reported here pertain to the moderate sample size of 100 patients. Due to the potential difficulty of no patients under observation for year 10, we collapsed the last two time intervals, and used nine instead of ten time intervals in the analysis, i.e., $(a_1, a_2, \ldots, a_K, a_{K+1}) = (0, 1, 2, 3, 4, 5, 6, 7, 8, 10)$ and $K = 9$.

We considered three censoring patterns: censoring at the ends of the intervals only, at the starts of the intervals only, and in the interiors of the intervals, which will be referred to as Cases I, II and III, respectively. The portions of the costs incurred after censoring times are, of course, unobservable. Recall that $\hat{E}_A$ and $\hat{E}_B$ are consistent in Cases I and II, respectively. Case II is the worst scenario for $\hat{E}_A$ since a censored patient then has observed cost of zero in the censoring interval. In real applications, one would move the a_k's slightly to the right so that $\hat{E}_A$ would be consistent. Similarly, one would not apply $\hat{E}_B$ literally to Case I. We included $\hat{E}_A$ in Case II and $\hat{E}_B$ in Case I (without any modifications) in order to assess the upper bounds of the biases for these two estimators.

We carried out two sets of simulation studies with different levels of censorship. In the first set, the censoring times have the following distributions: Case I: $\Pr(U = a_k-) = 0.05$ $(k = 2, \ldots, 10)$ and $\Pr(U = a_{10}) = 0.55$; Case II: $\Pr(U = a_k) = 0.05$ $(k = 1, \ldots, 9)$ and $\Pr(U = a_{10}) = 0.55$; Case III: $\Pr(U < t) = t/20$ $(t < a_{10})$ and $\Pr(U = a_{10}) = 0.50$. For the second set, we increased the amount of censoring as follows: Case I: $\Pr(U = a_k-) = 0.08$ $(k = 2, \ldots, 10)$ and $\Pr(U = a_{10}) = 0.28$; Case II: $\Pr(U = a_k) = 0.08$ $(k = 1, \ldots, 9)$ and $\Pr(U = a_{10}) = 0.28$; Case III: $\Pr(U < t) = t/12.5$ $(t < a_{10})$ and $\Pr(U = a_{10}) = 0.20$. The overall censoring probabilities for the two sets are approximately 25% and 40% under the uniform survival time distribution and about 30% and 45% under the exponential distribution. We shall refer to the first set as light censoring and to the second set as moderate censoring.

The main results of the simulation studies are summarized in Tables 1 and 2. We first comment on the performance of the methods developed in §2.2. The estimators $\hat{E}_A$ and $\hat{E}_B$ appear to be unbiased in Cases I and II, respectively. The variance (or standard error) estimators for $\hat{E}_A$ and $\hat{E}_B$ provide fairly accurate estimation of the true variation, and the confidence interval based on the normal approximation has adequate coverage probability when the bias of the estimator ($\hat{E}_A$ or $\hat{E}_B$) itself is small. There are considerable biases for $\hat{E}_A$ in Cases II and III especially with moderate censoring, which implies that narrower time intervals, quarterly or monthly, should be used for this estimator when censoring does not occur exclusively at the ends of the intervals. The estimator $\hat{E}_B$ performs well in all three cases. We recommend that $\hat{E}_A$ be used when censoring is concentrated at the ends of the intervals and $\hat{E}_B$ be used in all other situations.

The estimator $\hat{E}_T$ is virtually unbiased in Cases I and II, but has some bias in Case III especially under moderate censoring. The variance (or standard error) estimator is fairly reliable under light censoring, but underestimates the true variation under moderate censoring. The corresponding confidence interval has reasonable coverage probability except in Case III with moderate censoring. We conclude that the methodology of §2.3 works well when censoring occurs solely at the boundaries of the intervals provided that there are a few (say, 5 or more) observed deaths in each interval, but the results need be interpreted with caution when there is substantial censoring in the interior of an interval.

It is interesting to compare $\hat{E}_T$ versus $\hat{E}_A$ and $\hat{E}_B$. For light censoring, the performance of $\hat{E}_T$ is comparable to, in fact appears to be slightly better than, that of $\hat{E}_A$ and $\hat{E}_B$. Under the exponential survival distribution with moderate censoring, however, $\hat{E}_T$ is noticeably worse than $\hat{E}_A$ in Case I and worse than $\hat{E}_B$ in Cases II and III.

As expected, the naive estimators $\hat{E}_F$ and $\hat{E}_U$ are much worse than $\hat{E}_A$ and $\hat{E}_B$, respectively, in all cases. The negative bias of $\hat{E}_F$ is very alarming, especially under moderate censoring. Due to the positive correlation between the survival time and the total cost, $\hat{E}_U$ is also biased downwards. (Obviously, $\hat{E}_U$ will be biased upwards if the patients who die early tend to accrue higher costs than those who live longer.) The bias of $\hat{E}_{KM}$ can be substantial and of either direction.

4. A Real Example

We now use the proposed methodology to study the medical costs attributable to epithelial ovarian cancer among Medicare enrollees in the United States. The data base (Potosky et al., 1993) consists of 4798 Medicare beneficiaries over the age of 65 who were diagnosed with local, regional or distant stage ovarian cancer from 1973 through 1989. The data collection was initiated at the beginning of 1984 and terminated at the end of 1990. As a result, the patients who died before 1984 were excluded and those still alive at the end of 1990 were censored. The cost information covers the years 1984 through 1990 and contains the monthly costs on those who were alive at some point during this period. Even though no individual patients were followed for more than seven years, the post-diagnosis information on survival and cost extended over a period of seventeen years (after diagnosis) due to staggered times of diagnosis.

In this study, the assumption of independent censoring is satisfied because study termination was the only source of censoring. Due to the lack of information on the exact date of diagnosis, we assume that the diagnosis took place at the start of the calendar month. This assumption, together with the use of December 31, 1990 as the censoring date, creates a situation where censoring occurred only at the end of each month. Thus, the estimators $\hat{E}_A$ and $\hat{E}_T$ with monthly intervals will be consistent.

The survival times were also subject to left-truncation in that the patients who were diagnosed before 1984 were not followed from their times of diagnosis. Because the truncation date is independent of the survival time and costs in this case, the methodology presented in §2.2 can be applied with some minor modifications, as explained at the end of Appendix A. The left-truncation and right-censorship generally do not affect the time period over which the average total cost may be estimated by this approach because it requires some cost data in each time interval, say each month, but not the total cost over the entire time period of interest on any particular patient. The Kaplan-Meier estimators involved in the construction of $\hat{E}_T$ may also be modified to accommodate left-truncation; see the end of Appendix B. However, only the patients whose total costs were fully observed may be used in the estimation of the A_k's. As a result, the methodology of §2.3 cannot be used to estimate the total cost over a time period longer than the maximal length of follow-up on a patient, which is seven years in this case.

Out of the 4789 patients, 1080, 1020 and 2698 were diagnosed as local, regional and distant stages, respectively. We first use the methodology of §2.2 to estimate the average 15-year post-diagnosis costs for the three clinical stages. The starting time for our analysis is taken to be one month prior to diagnosis so as to incorporate the costs associated with making the definite diagnosis.

The Kaplan-Meier estimates of the survival probabilities are displayed in Figure 1, and the cost estimates based on three methods are summarized in Table 3. As discussed earlier, $\hat{E}_A$ is consistent in this case. Due to heavy left-truncation and right-censorship, the naive full-sample estimate $\hat{E}_F$ substantially underestimates the true cost, especially for local and regional stages. The naive complete-cases estimator $\hat{E}_U$, which excludes the truncated and censored cases, is also severely biased downwards as none of the complete cases were followed for more than seven years. The naive Kaplan-Meier estimator $\hat{E}_{KM}$ cannot even be calculated in this case because the costs accrued prior to the truncation times are unknown.

The 95% confidence intervals based on $\hat{E}_A$ are $(3273, 10636)$ and $(1287, 6843)$ for comparing local vs. regional and regional vs. distant stages, respectively, indicating that the long-term costs are the highest for those diagnosed with local stage, and the lowest for the distant stage. The cost histories are displayed in Figure 2. This figure, combined with Figure 1, shows that, although it is highly expensive to treat late stage ovarian cancer, the lifetime costs are higher among patients diagnosed at a less advanced stage because those patients have longer post-diagnosis survival times.

For further illustration, we use the methodology of §2.3 to analyze the seven-year cost. Because most of the patients diagnosed with distant-stage ovarian cancer died within seven years after diagnosis whereas the regional-stage and especially the local-stage patients lived considerably longer, it is more interesting to apply the methodology to the former patients than to the latter patients. We henceforth concentrate on estimating the average cost during the first seven years for the patients diagnosed with distant-stage ovarian cancer.

Approximately 71.2% of the 2698 patients diagnosed with distant-stage ovarian cancer were followed from diagnosis to death. As shown in Table 4, $\hat{E}_T$ produces a slightly lower estimate of the 7-year average cost than $\hat{E}_A$. Neither estimate is far away from the 15-year estimate of $\hat{E}_A$ given in Table 3 because the distant-stage patients rarely lived past the seven-year mark. The standard error estimate for $\hat{E}_T$ is about 13% higher than that of $\hat{E}_A$, which entails that the confidence interval based on $\hat{E}_T$ is a bit wider than that of $\hat{E}_A$. As expected, $\hat{E}_F$ and $\hat{E}_U$ give smaller estimates of the average 7-year cost.

Figure 3 displays the estimated distribution function for the 7-year cost along with the pointwise 95% confidence intervals. The distribution is highly skewed in that most (about 73%) of the seven-year costs are less than \$50000 whereas some patients accrue costs of more than \$200000. The estimation of the median cost is also illustrated in the figure. The point estimate is \$29935, which is about \$8000 less than the mean estimate shown in Table 4. The approximate 95% confidence interval for the median cost is (\$28375, \$31379).

5. Discussions

This paper provides a rigorous treatment of the important problem of cost estimation in medical studies. The proposed estimators are consistent under appropriate censoring conditions and are asymptotically normal with easily estimated variances. The numerical results in the previous two sections demonstrated that these estimators perform well in practical settings. By contrast, the commonly used naive sample averages can be very misleading.

The methodology described in §2.2 requires that the cost histories be recorded on some patients whereas that of §2.3 requires only the total costs at the last contact dates (among those who are observed to die before τ or still alive at τ). If the cost histories are available, then the former approach is usually preferable, especially when there is substantial censoring and/or truncation, as it makes fuller use of the cost information and requires smaller sample sizes. However, only the latter approach can be used to estimate the distribution function and quantiles.

The cost information and survival information need not come from the same set of patients or the same source of data. We may accommodate such situations by making some minor changes in our notation. Suppose that there is a total of n patients, some of whom are used to estimate the E_k's (or the A_k's) and some of whom are used to estimate the S_k's. If the ith patient is not involved in estimating E_k (or A_k), then we set $Y_{ki} = 0$. Similarly, if the jth patient is not involved in estimating S_k $(k = 1, \ldots, K)$, then we set $X_j = \delta_j = 0$. With these modifications, all the results given in §2 and the Appendices, including the variance formulas, will hold.

The $\hat{E}_A$ estimates of the average costs shown in Table 3 differ slightly from those of Etzioni et al. (1995), who used a different database for the survival estimation in order to bypass the problem of left-truncation. It would not be possible to estimate the variances for their estimates because some patients belong to both data sources but we do not have the information to match those patients.

Our definition of the average cost E includes both the patients who die before τ and those who are still alive at τ. Thus, $\hat{E}_T$ requires the observed costs over $[0, \tau)$ on those still alive at τ. An alternative definition is to exclude the latter patients. Then E would be interpreted as the average total cost among those who die in $[0, \tau)$ rather than the average total cost over $[0, \tau)$ among all patients. Under the alternative definition, it would be possible to estimate the average cost over some sub-interval of $[0, \tau)$ in the absence of the cost history data.

The assumption of independent censoring requires some care. This assumption is clearly not satisfied if patients are withdrawn from the study for health- or cost-related reasons. It is very difficult, if not impossible, to deal with such dependent censoring even for the survival time distribution itself. One must carefully examine the independent censoring assumption before applying the proposed methodology.

Acknowledgements

The authors are grateful to Paula Diehr, Scott Emerson, Sean Sullivan and Nicole Urban for useful discussions and to Arnie Potosky and Gerald Riley for accessing the linked SEER-Medicare database. This research was supported by the NIH grants R29 GM-47845 and R01 AI-291968 (for Lin) and the DOD grant DAMD 17-94-J-4237 (for Etzioni).

REFERENCES

Brookmeyer, R. and Crowley, J. (1982). A confidence interval for the median survival time. *Biometrics*, **38**, 29–41.

Etzioni, R., Urban, N. and Baker, M. (1995). Estimating the costs attributable to a disease with application to ovarian cancer. *J. Clin. Epi.*, in press.

Fenn, P., et al. (1995). The analysis of censored treatment cost data in economic evaluation. *Medical Care*, in press.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Hiatt, R. A., et al. (1990). The cost of acquired immunodeficiency syndrome in Northern California: the experience of a large prepaid health plan. *Arch. Intern. Med.*, **150**, 833–838.

Hodgson, T. A. (1992). Cigarette smoking and lifetime medical expenditures. *The Milbank Quarterly*, **70**, 81–125.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Keeler, E. B., et al. (1989). The external costs of a sedentary lifestyle. *Am. J. Public Health*, **79**, 975–981.

Manning, W. G., et al. (1989). The taxes of sin. Do smokers and drinkers pay their way? *J. Am. Med. Ass.*, **261**, 1604–1609.

Potosky, A. L., et al. (1993). Potential for cancer related health services research using a linked Medicare-tumor registry data base. *Medical Care*, **31**, 732–747.

Quesenberry, C. P., et al. (1989). A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome. *Am. J. Public Health*, **79**, 1643–1647.

APPENDIX A
Large-Sample Properties of $\hat{E}_A$ and $\hat{E}_B$

By the law of large numbers, the estimators $\hat{E}_k$ $(k = 1, \ldots, K)$ converge in probability to E_k^*. It then follows from Slutsky's theorem and the consistency of the Kaplan-Meier estimator that $\hat{E}$ ($\hat{E}_A$ or $\hat{E}_B$) converges in probability to E^*.

Let $Z = n^{1/2}(\hat{E} - E^*)$. Then

$$\begin{aligned} Z &= n^{1/2}\left(\sum_{k=1}^{K} \hat{S}_k \hat{E}_k - \sum_{k=1}^{K} S_k E_k^*\right) \\ &= n^{1/2} \sum_{k=1}^{K} \hat{S}_k(\hat{E}_k - E_k^*) + n^{1/2} \sum_{k=1}^{K} E_k^*(\hat{S}_k - S_k) \\ &= Z_1 + Z_2, \quad \text{say.} \end{aligned}$$

Due to the consistency of the $\hat{S}_k$'s,

$$Z_1 = n^{1/2} \sum_{k=1}^{K} S_k(\hat{E}_k - E_k^*) + o_p(1) = \sum_{k=1}^{K} S_k \frac{n^{-1/2} \sum_{i=1}^{n} Y_{ki}(\tilde{C}_{ki} - E_k^*)}{n^{-1} \sum_{i=1}^{n} Y_{ki}} + o_p(1),$$

where $o_p(1)$ denotes an asymptotically negligible term, which converges in probability to 0. By the central limit theorem and the law of large numbers, the random variable $n^{-1/2}\sum_{i=1}^{n} Y_{ki}(\tilde{C}_{ki} - E_k^*)$ is asymptotically zero-mean normal and the random variable $n^{-1}\sum_{i=1}^{n} Y_{ki}$ converges in probability to the constant $\mathcal{E}(Y_{k1})$. It then follows from Slutsky's theorem that

$$Z_1 = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{S_k Y_{ki}(\tilde{C}_{ki} - E_k^*)}{\mathcal{E}(Y_{k1})} + o_p(1), \tag{A.1}$$

which is essentially a sum of n i.i.d. (independent and identically distributed) zero-mean random variables.

We shall also derive an i.i.d. representation for Z_2. It is convenient to introduce the counting processes $N_i(t) = \delta_i I(X_i \leq t)$ and the associated martingales $M_i(t) = N_i(t) - \int_0^t I(X_i \geq t)d\Lambda(t)$ $(i = 1, \ldots, n)$, where $\Lambda(\cdot)$ is the cumulative hazard function of the survival time T. It is well-known that the Kaplan-Meier estimator $\hat{S}_k$ is asymptotically equivalent to $e^{-\hat{\Lambda}_k}$, where $\hat{\Lambda}_k$ is the Nelson-Aalen estimator for $\Lambda_k = \Lambda(a_k)$, i.e.,

$$\hat{\Lambda}_k = \sum_{i=1}^{n}\int_0^{a_k}\frac{dN_i(t)}{\sum_{j=1}^{n} I(X_j \geq t)}.$$

Furthermore,

$$n^{1/2}(\hat{\Lambda}_k - \Lambda_k) = n^{-1/2}\sum_{i=1}^{n}\int_0^{a_k}\frac{dM_i(t)}{n^{-1}\sum_{j=1}^{n} I(X_j \geq t)} + o_p(1) \tag{A.2}$$

(Fleming and Harrington, 1991, pp. 5-6). Then the martingale central limit theorem (Fleming and Harrington, 1991, Theorem 5.3.5) enables us to replace the denominator on the right side of (A.2) by its expectation, yielding

$$n^{1/2}(\hat{\Lambda}_k - \Lambda_k) = n^{-1/2}\sum_{i=1}^{n}\int_0^{a_k}\frac{dM_i(t)}{\Pr(X \geq t)} + o_p(1),$$

which is essentially a sum of n i.i.d. zero-mean random variables. By the Taylor series expansion, $n^{1/2}(\hat{S}_k - S_k) = -S_k n^{1/2}(\hat{\Lambda}_k - \Lambda_k) + o_p(1)$. Thus,

$$Z_2 = -n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} E_k^* S_k \int_0^{a_k}\frac{dM_i(t)}{\Pr(X \geq t)} + o_p(1). \tag{A.3}$$

Combination of equations (A.1) and (A.3) yields $Z = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{ki} + o_p(1)$, where

$$\xi_{ki} = \frac{S_k Y_{ki}(\tilde{C}_{ki} - E_k^*)}{\mathcal{E}(Y_{k1})} - E_k^* S_k \int_0^{a_k}\frac{dM_i(t)}{\Pr(X \geq t)}. \tag{A.4}$$

Note that, for every i, the random elements involved in ξ_{ki} ($k = 1, \ldots, K$) pertain to the ith patient only, which indicates that the random variable Z is essentially a sum of n i.i.d. zero-mean random variables. Therefore, a straightforward application of the central limit theorem shows that Z converges in distribution to a zero-mean normal random variable with variance $\sigma^2 = \mathcal{E}\left(\sum_{k=1}^K \sum_{l=1}^K \xi_{k1}\xi_{l1}\right)$.

It is natural to estimate σ^2 by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K \hat{\xi}_{ki}\hat{\xi}_{li}$, where the $\hat{\xi}_{ki}$'s are obtained from the ξ_{ki}'s by replacing the unknown quantities on the right side of (A.4) with their respective sample estimators, i.e.,

$$\hat{\xi}_{ki} = \frac{\hat{S}_k Y_{ki}(\tilde{C}_{ki} - \hat{E}_k)}{n^{-1}\sum_{j=1}^n Y_{kj}} - \hat{E}_k \hat{S}_k \int_0^{a_k} \frac{dN_i(t) - I(X_i \geq t) d\hat{\Lambda}(t)}{n^{-1}\sum_{l=1}^n I(X_l \geq t)}.$$

The integral on the right side of the above equation equals

$$\frac{I(X_i \leq a_k)\delta_i}{n^{-1}\sum_{l=1}^n I(X_l \geq X_i)} - \sum_{j=1}^n \frac{I(X_i \geq X_j) I(X_j \leq a_k)\delta_j}{n^{-1}\{\sum_{l=1}^n I(X_l \geq X_j)\}^2}. \qquad (A.5)$$

Note that the variance expression $\hat{V}$ for $\hat{E}$ given in §2.2 is simply $\hat{\sigma}^2/n$. The consistency of $\hat{\sigma}^2$ follows from the consistency of $\hat{\Lambda}(\cdot)$, $\hat{E}_k$ and $\hat{S}_k$ ($k = 1, \ldots, K$) along with Slutsky's theorem and the law of large numbers.

To provide some insights into the limiting variance expression, let $\xi_{ki}^{(1)}$ and $\xi_{ki}^{(2)}$ denote the two terms on the right side of (A.4), i.e., $\xi_{ki} = \xi_{ki}^{(1)} + \xi_{ki}^{(2)}$. Then

$$\sigma^2 = \mathcal{E}\left(\sum_{k=1}^K \sum_{l=1}^K \xi_{k1}^{(1)}\xi_{l1}^{(1)}\right) + \mathcal{E}\left(\sum_{k=1}^K \sum_{l=1}^K \xi_{k1}^{(2)}\xi_{l1}^{(2)}\right) + 2\mathcal{E}\left(\sum_{k=1}^K \sum_{l=1}^K \xi_{k1}^{(1)}\xi_{l1}^{(2)}\right), \qquad (A.6)$$

which is just a representation for $\text{Var}(Z) = \text{Var}(Z_1) + \text{Var}(Z_2) + 2\text{Cov}(Z_1, Z_2)$. The first two terms on the right side of (A.6) are the variances attributable to the variations of the $\hat{E}_k$'s and the $\hat{S}_k$'s, respectively, and the third term is the covariance. Each of the three terms account for the variations within the intervals as well as the covariances among the intervals.

Finally, we show how to incorporate left-truncation into our estimation procedures. Let B_i ($i = 1, \ldots, n$) be the truncation times, which are assumed to be independent of the survival times and costs. Since the costs incurred before the truncation times are not recorded, we set Y_{ki} to be 0 if $B_i > a_k$, or include the indicator $I(B_i \leq a_k)$ in the definition of Y_{ki}. The risk sets for the survival estimation are modified in a similar fashion. Specifically, expression (A.5) becomes

$$\frac{I(X_i \leq a_k)\delta_i}{n^{-1}\sum_{l=1}^n I(B_l \leq X_i \leq X_l)} - \sum_{j=1}^n \frac{I(B_i \leq X_j \leq X_i) I(X_j \leq a_k)\delta_j}{n^{-1}\{\sum_{l=1}^n I(B_l \leq X_j \leq X_l)\}^2}.$$

APPENDIX B

Large-Sample Properties of $\hat{E}_T$

By the law of large numbers, the estimators $\hat{A}_k$ $(k = 1, \ldots, K+1)$ converge in probability to A_k^*. It then follows from Slutsky's theorem and the consistency of the Kaplan-Meier estimator that $\hat{E}_T$ converges in probability to E^*, which reduces to E if censoring occurs only at the boundaries of the intervals.

Write $Z = n^{1/2}(\hat{E} - E^*)$. We decompose Z as $Z = Z_1 + Z_2 - Z_3$, where

$$Z_1 = n^{1/2} \sum_{k=1}^{K+1} (\hat{S}_k - \hat{S}_{k+1})(\hat{A}_k - A_k^*),$$

$$Z_2 = n^{1/2} \sum_{k=1}^{K+1} A_k^*(\hat{S}_k - S_k),$$

$$Z_3 = n^{1/2} \sum_{k=1}^{K+1} A_k^*(\hat{S}_{k+1} - S_{k+1}).$$

Because of the consistency of the $\hat{S}_k$'s,

$$Z_1 = n^{1/2} \sum_{k=1}^{K+1} (S_k - S_{k+1})(\hat{A}_k - A_k^*) + o_p(1) = \sum_{k=1}^{K+1} (S_k - S_{k+1}) \frac{n^{-1/2} \sum_{i=1}^n Y_{ki}(C_i - A_k^*)}{n^{-1} \sum_{i=1}^n Y_{ki}} + o_p(1).$$

By the arguments given in Appendix A, one may replace $n^{-1} \sum_{i=1}^n Y_{ki}$ in the above equation by its limit $\mathcal{E}(Y_{k1})$ without altering the asymptotic distribution of Z_1. This replacement yields the following i.i.d. representation

$$Z_1 = n^{-1/2} \sum_{i=1}^n \sum_{k=1}^{K+1} \frac{(S_k - S_{k+1}) Y_{ki}(C_i - A_k^*)}{\mathcal{E}(Y_{k1})} + o_p(1). \qquad (B.1)$$

It also follows from the arguments of Appendix A that

$$Z_2 = -n^{-1/2} \sum_{i=1}^n \sum_{k=1}^{K+1} A_k^* S_k \int_0^{a_k} \frac{dM_i(t)}{\Pr(X \geq t)} + o_p(1), \qquad (B.2)$$

and

$$Z_3 = -n^{-1/2} \sum_{i=1}^n \sum_{k=1}^{K+1} A_k^* S_{k+1} \int_0^{a_{k+1}} \frac{dM_i(t)}{\Pr(X \geq t)} + o_p(1), \qquad (B.3)$$

where $M_i(t)$ $(i = 1, \ldots, n)$ are the martingales introduced in Appendix A.

By combining equations (B.1)–(B.3), we obtain $Z = n^{-1/2} \sum_{i=1}^n \sum_{k=1}^{K+1} \xi_{ki} + o_p(1)$, where

$$\xi_{ki} = \frac{(S_k - S_{k+1}) Y_{ki}(C_i - A_k^*)}{\mathcal{E}(Y_{k1})} + A_k^* \left\{ S_{k+1} \int_0^{a_{k+1}} \frac{dM_i(t)}{\Pr(X \geq t)} - S_k \int_0^{a_k} \frac{dM_i(t)}{\Pr(X \geq t)} \right\}. \qquad (B.4)$$

Because Z is essentially a sum of n i.i.d. zero-mean random variables, it follows from the central limit theorem that Z converges in distribution to a zero-mean normal random variable with variance $\sigma^2 = \mathcal{E}\left(\sum_{k=1}^{K+1}\sum_{l=1}^{K+1}\xi_{k1}\xi_{l1}\right)$. As argued in Appendix A, the limiting variance σ^2 can be consistently estimated by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K+1}\sum_{l=1}^{K+1}\hat{\xi}_{ki}\hat{\xi}_{li}$, where the $\hat{\xi}_{ki}$'s are obtained from the ξ_{ki}'s by replacing the unknown quantities on the right side of (B.4) with their respective sample estimators, i.e.,

$$\hat{\xi}_{ki} = \frac{n(\hat{S}_k - \hat{S}_{k+1})Y_{ki}(C_i - \hat{A}_k)}{\sum_{j=1}^{n} Y_{kj}} + n\hat{A}_k(\hat{S}_{k+1}D_{k+1,i} - \hat{S}_k D_{ki}),$$

where D_{ki} is defined by (2.4). (Note that $W_{ki} = \hat{\xi}_{ki}/n$.)

If there is left-truncation, then the risk sets involved in the Kaplan-Meier estimation will be adjusted accordingly, the truncated cases will be excluded from the calculation of the $\hat{A}_k$'s, and D_{ki} will be changed to

$$D_{ki} = \frac{I(X_i \le a_k)\delta_i}{\sum_{l=1}^{n} I(B_l \le X_i \le X_l)} - \sum_{j=1}^{n} \frac{I(B_i \le X_j \le X_i)I(X_j \le a_k)\delta_j}{\{\sum_{l=1}^{n} I(B_l \le X_j \le X_l)\}^2},$$

where B_i $(i = 1, \ldots, n)$ are the truncation times, which are assumed to be independent of the survival times and costs.

Table 1.

Summary statistics for the simulation studies: light censoring

Estimator		Uniform survival times			Exponential survival times		
		Case I	Case II	Case III	Case I	Case II	Case III
$\hat{E}_A$	Bias	–4	–1837	–986	–2	–1503	–819
	SSE	1148	1179	1152	1139	1139	1129
	SEE	1116	1147	1119	1115	1120	1109
	CP	94.1%	64.0%	84.7%	94.3%	72.1%	87.2%
$\hat{E}_B$	Bias	279	–4	–29	324	–1	86
	SSE	1112	1190	1133	1149	1177	1161
	SEE	1080	1152	1097	1127	1152	1136
	CP	93.2%	94.0%	94.0%	93.6%	94.2%	94.2%
$\hat{E}_T$	Bias	–3	–4	–48	–1	–1	–24
	SSE	1112	1149	1144	1141	1175	1170
	SEE	1093	1127	1113	1096	1126	1119
	CP	94.3%	94.2%	94.0%	93.7%	93.7%	93.6%
$\hat{E}_F$	Bias	–5418	–6865	–6180	–3877	–5109	–4528
	SSE	1259	1333	1292	1149	1208	1174
	SEE	1252	1326	1284	1134	1196	1161
	CP	0.8%	0.0%	0.2%	7.4%	1.2%	2.8%
$\hat{E}_U$	Bias	–1283	–1373	–1423	–468	–544	–572
	SSE	1187	1228	1211	1318	1357	1338
	SEE	1176	1216	1199	1295	1335	1316
	CP	81.1%	80.2%	78.5%	93.0%	92.7%	92.4%
$\hat{E}_{KM}$	Bias	–832	–877	–880	1135	1096	1062
	SSE	1124	1158	1142	1181	1211	1195
	SEE	1103	1134	1119	1147	1175	1160
	CP	88.0%	87.5%	87.2%	82.0%	83.5%	83.9%

Note: The true mean costs are 39000 and 34680 dollars under the uniform and exponential distributions, respectively. Bias and SSE are, respectively, the sampling bias and sampling standard error for the estimator. SEE is the sampling average of the standard error estimator, and CP is the sampling coverage probability of the 95% confidence interval. Cases I, II and III correspond to censoring at the ends, at the starts and in the interiors of the intervals, respectively. Each entry is based on 50000 simulation samples.

Table 2.

Summary statistics for the simulation studies: moderate censoring

		Uniform survival times			Exponential survival times		
Estimator		Case I	Case II	Case III	Case I	Case II	Case III
$\hat{E}_A$	Bias	−1	−3692	−2032	−1	−2920	−1652
	SSE	1304	1364	1303	1287	1276	1258
	SEE	1248	1317	1253	1247	1243	1225
	CP	93.7%	21.5%	62.3%	93.7%	36.5%	70.4%
$\hat{E}_B$	Bias	546	−9	−156	679	−4	214
	SSE	1225	1423	1290	1358	1408	1433
	SEE	1178	1344	1221	1304	1345	1337
	CP	91.5%	93.3%	92.8%	91.4%	93.1%	92.4%
$\hat{E}_T$	Bias	−7	−17	−317	−11	−3	−93
	SSE	1221	1339	1537	1326	1431	1530
	SEE	1183	1262	1263	1218	1281	1283
	CP	94.0%	93.5%	90.5%	92.3%	91.4%	90.2%
$\hat{E}_F$	Bias	−8663	−10983	−9885	−6201	−8174	−7244
	SSE	1277	1331	1296	1159	1213	1178
	SEE	1267	1320	1286	1144	1198	1163
	CP	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
$\hat{E}_U$	Bias	−2470	−2819	−2845	−1422	−1677	−1670
	SSE	1311	1400	1358	1389	1463	1424
	SEE	1295	1381	1340	1363	1437	1400
	CP	52.5%	47.1%	43.8%	81.9%	78.6%	77.9%
$\hat{E}_{KM}$	Bias	−1572	−1769	−1722	404	234	222
	SSE	1195	1257	1224	1220	1276	1242
	SEE	1165	1224	1193	1180	1230	1200
	CP	72.4%	69.1%	69.2%	92.7%	93.4%	93.5%

Note: See the Note to Table 1.

Table 3.

Estimates and 95% confidence intervals for the average 15-year costs attributable to ovarian cancer separated by clinical stages at diagnosis

Method	Local stage	Regional stage	Distant stage
$\hat{E}_A$	50620 (47891, 53349)	43666 (41195, 46137)	39601 (38331, 40870)
$\hat{E}_F$	26390 (24884, 27895)	34706 (32915, 36497)	33746 (32698, 34793)
$\hat{E}_U$	44350 (39402, 49297)	38092 (35675, 40508)	35221 (33965, 36478)

Table 4.

Estimation of the 7-year average cost for distant-stage ovarian cancer patients

Method	Estimate	Stand. Error	95% Confidence Interval
$\hat{E}_T$	37544	702	(36168, 38920)
$\hat{E}_A$	38452	621	(37236, 39668)
$\hat{E}_F$	33775	534	(32728, 34822)
$\hat{E}_U$	35452	638	(34202, 36702)

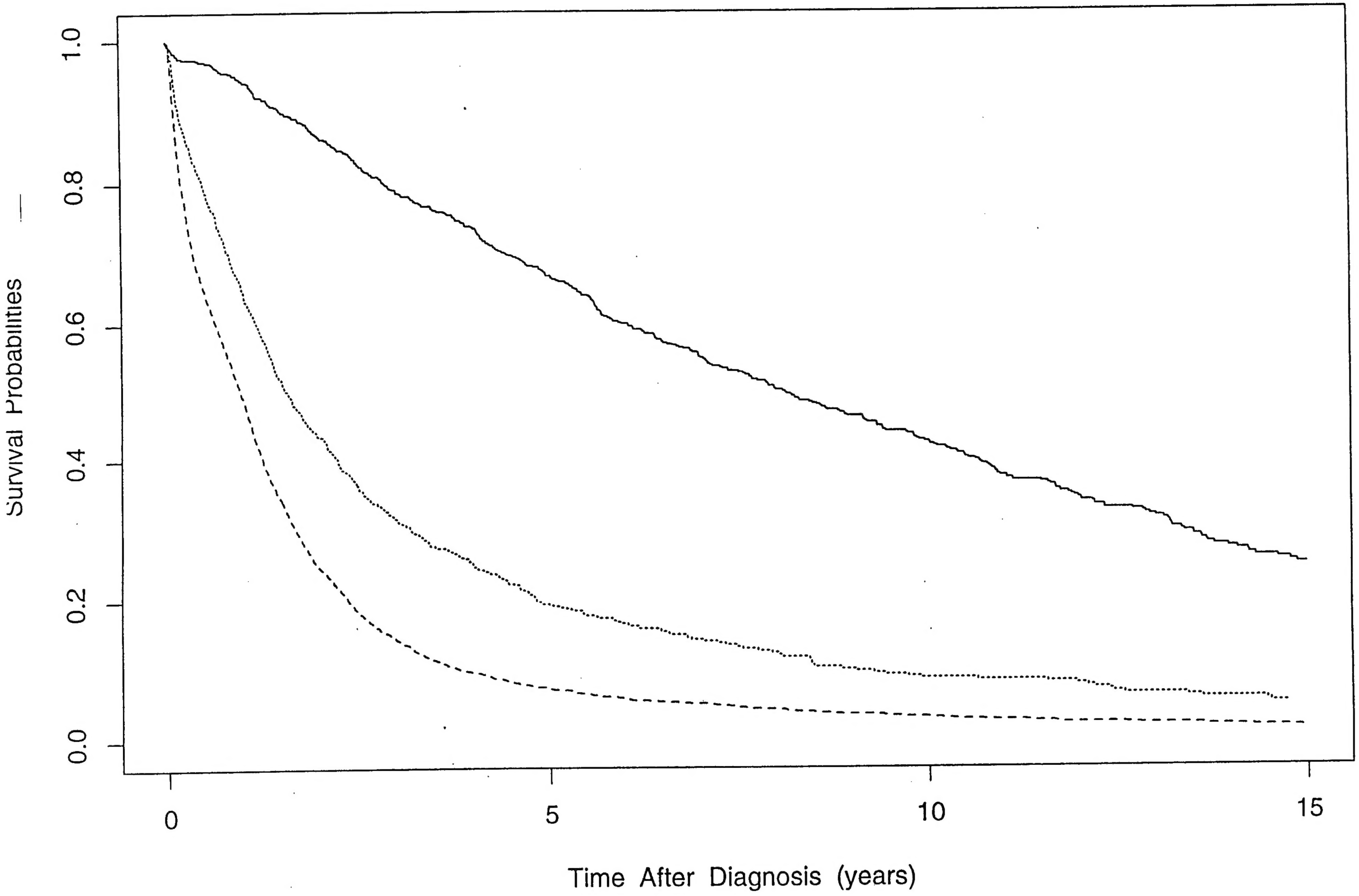
Figure Legends

Figure 1. Kaplan-Meier estimates of the survival probabilities for epithelial ovarian cancer patients: local stage, shown by solid curve; regional stage, shown by dotted curve; distant stage, shown by dashed curve.

Figure 2. Estimates of the average cumulative costs for epithelial ovarian cancer patients: local stage, shown by solid curve; regional stage, shown by dotted curve; distant stage, shown by dashed curve.

Figure 3. Estimation of the distribution function of the 7-year cost for distant-stage ovarian cancer patients: point estimate, shown by the middle solid curve; pointwise 95% confidence intervals, the dashed curves. The point estimate and the 95% confidence interval for the median cost are the intercepts of the vertical lines with the horizontal axis.

Figure 1



Survival Probabilities —
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Time After Diagnosis (years)

Figure 2

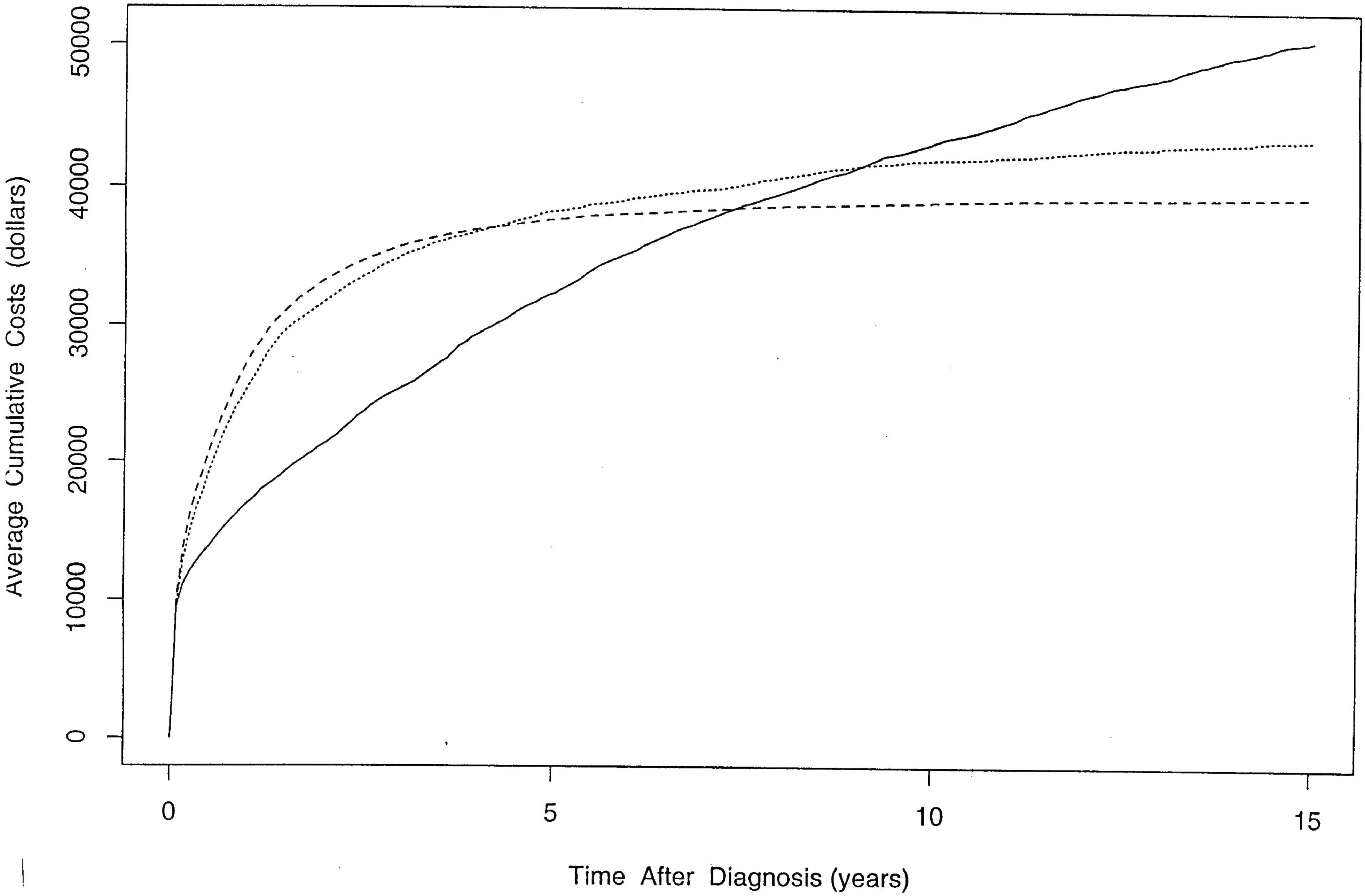
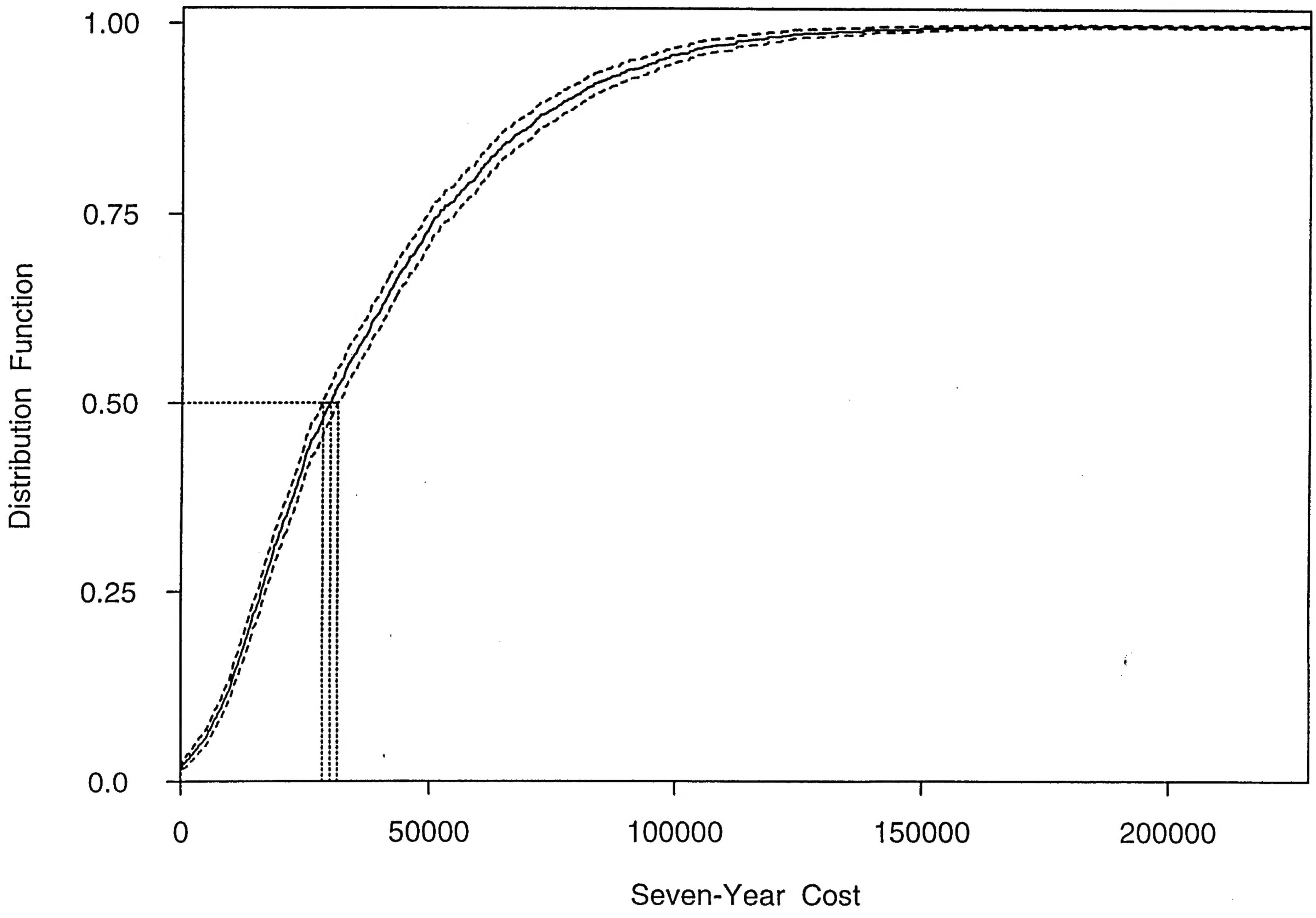

Average Cumulative Costs (dollars)
0
10000
20000
30000
40000
50000
0
5
10
15
Time After Diagnosis (years)

Figure 3


1.00
0.75
0.50
0.25
0.0
Distribution Function
0
50000
100000
150000
200000
Seven-Year Cost